

PROBIT AND LOGIT ANALYSES

by

JANET LEE DUNCAN

B. S., Kansas State University, 1963

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

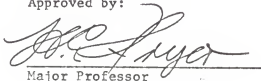
MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by:


Major Professor

10
2008
8.1
9.1
D9
r12

CONTENTS

0. INTRODUCTION.....	1
1. GENERAL STATISTICAL MODEL.....	2
2. PROBIT ANALYSIS.....	4
3. LOGIT ANALYSIS.....	16
3.1 THE LOGISTIC CURVE.....	18
3.2 APPLYING THE LOGISTIC TO BIOASSAY.....	21
4. CONCLUSION.....	28
REFERENCES.....	30
ACKNOWLEDGEMENTS.....	32

O. INTRODUCTION

Probit analysis and logit analysis are used in the general area of biological assay. In bioassay, plants or animals are subjected to some stimulus in varying intensities and the response of the organism to each intensity is observed. The relationship between the intensities of the stimulus and the responses they elicit is then inferred from the observations. The stimulus can cover a wide range of chemical, physical, biological, physiological, or psychological agents which produce an observable response when administered to a particular organism.

Biological assays usually have one of two purposes:

1. To determine the mathematical relationship between the intensity of the stimulus and the level of the response.
2. To evaluate the unknown strength of an agent by observing the response it elicits in organisms whose response relationship with an agent of known strength has been determined previously.

When response is known as a function of the stimulus, predictions can be made of intensities of the stimulus which will produce desirable responses. For example, after a certain concentration of insecticide is reached, higher concentrations produce almost no additional increase in death rate and the additional expense of further concentration could be avoided. To illustrate the second type of assay, suppose a new method of manufacturing the insecticide has been discovered but its strength is unknown. To assay its strength, the mortality rate it produces is observed in insects for which the functional relationship between strength of the insecticide prepared by the old method and mortality rate is known. Response is assumed to be independent of the method of preparation and

dependent only on the active ingredient present. A fifty percent kill under the unknown strength of the new method would then be equated with the strength of the old preparation that also produced a fifty percent kill.

This report will deal with the particular type of bioassay in which the stimulus is the dose of a toxic agent and the response observed is quantal. Quantal responses are those in which the proportion of organisms affected by a particular dosage is observed out of the total number of organisms exposed, as opposed to a response measured on a continuous scale, such as weight or length. The all-or-none responses of death or survival given in the insecticide example are quantal responses.

Because both of the purposes of bioassay involve finding the mathematical relationship between dosage of an agent and response of an organism, the main procedure is one of curve-fitting. A curve is fitted to the observed data of an experiment to discover the mathematical relationship assumed to exist between agent and organism, and to minimize the deviations from it. These deviations will then be attributed to sampling error.

1. GENERAL STATISTICAL MODEL

The following definitions will be used in developing the general statistical model for quantal response

1. d_i = dosage of intensity i .
2. $x_i = \log_{10} d_i$
3. n_i = total number of organisms exposed to a given dose d_i of a toxic agent.
4. r_i = observed number of organisms responding to a dose d_i where response usually means death.

$$5. \quad p_i = \frac{r_i}{n_i} = \text{observed proportion of organisms responding to dose } d_i$$

out of the n_i organisms exposed; i.e., observed mortality rate.

$$6. \quad q_i = 1 - p_i$$

$$7. \quad P_i = \text{true mortality rate at } d_i$$

$$8. \quad Q_i = 1 - P_i$$

For convenience the subscript i will be omitted in the discussion when the meaning is clear.

In the statistical model for quantal response it is assumed that the observed response at a given dosage is distributed about the true response at that dosage, a binomial with mean P and variance PQ/n . Thus for a sample of n organisms acting independently of one another at a given dose D , the observed number responding, r , is binomially distributed. Then the probability that r individuals respond out of n exposed is given by:

$$(1) \quad \binom{n}{r} p^r q^{n-r}.$$

Dosage-mortality studies have been made upon a large variety of organisms by many biologists. These studies have established that a graph showing the percentage of dead organisms as the ordinate against some function of dosage (usually the log dose) as the abscissa is generally sigmoidal. The rate of change in percent kill per unit of dose is the lowest when the mortality rate is near zero and one hundred percent and is the highest at mortality rates near fifty percent. (It should be pointed out that a dosage, D , or log dose, x , could refer to exposure time of a fixed amount of stimulus, such as exposure time to X-rays.)

Among multicellular organisms it is practically universal for a graph of dosage versus mortality rate to result in a characteristic sigmoidal

shape; but there is more than one interpretation of this curve. It is these interpretations, or the underlying assumptions of the processes involved in the dosage-response relationship, that lead to the different methods of expressing the relationship mathematically. Both of the methods to be considered herein begin with a transformation which will rectify (make linear) the sigmoidal curve. The rectification will allow the use of linear regression analysis.

The most well-known of these methods, indeed apparently often considered synonymous with bioassay, is known as probit analysis. Probit analysis was essentially suggested by Gaddum, refined by R. A. Fisher, and actively promoted by C. I. Bliss.

The second method, devised by Joseph Berkson, from earlier work done by Pearl and Reed, and Wilson and Worchester, is called logit analysis.

From the purely empirical viewpoint of curve-fitting described previously, both the use of probits and the use of logits usually lead to essentially the same rectification of the originally curvilinear data, and to essentially the same fitted straight line. They differ basically in the underlying assumptions in what statistical tests might be run, and in the ease of the calculations required to fit a dosage-mortality curve.

Since probits were historically first and are by far the most widely used, they will be discussed first.

2. PROBIT ANALYSIS

The basic assumption of probit analysis is that the survival of an organism in the presence of a dose of a toxic agent is proportional to its tolerance to that agent. This assumption is widely accepted and seems to include the many complicated physical and chemical reactions which may occur

between the organism and the agent. As stated by Finney, "for quantal response it is necessary to consider the distribution of tolerances over the population studied." On the basis of the assumption of the existence of tolerances, the dosage-mortality curve is taken to be primarily a description of the variation in susceptibility between individuals of the population. If the susceptibility of an individual is represented by the smallest dose that is just sufficient to kill it, the number of individuals having each particular susceptibility might be expected to be normally distributed with respect to some function of dosage.

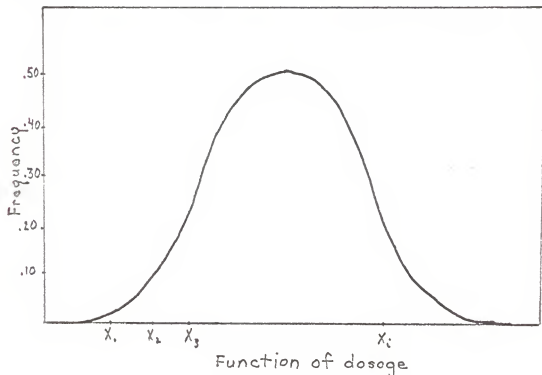


Fig. 1. Ordinates give the percentage of organisms in a single sample responding to an individual lethal dose X .

The exact lethal dose for each individual would be necessary to plot a normal frequency distribution of tolerances from sample data. However, experimental techniques with toxic agents are usually not sufficiently refined to enable the exact individual lethal dose for complicated organism to be

distinguished. Because all those organisms that succumb to a lower dose should also succumb to a higher dose, the observed mortality consists of all individuals who are susceptible to any dose between zero and the administered dose. Thus, the proportion of the total population responding to a dose X_0 is given by

$$P = \int_0^{X_0} f(X) dX.$$

If these percentage kills were then plotted as the ordinate of a new graph against the same function of dosage as the abscissa, the result should be the cumulative normal distribution function. As was noted previously, this is the general form of the curve obtained when plotting percentage killed against the log dose.

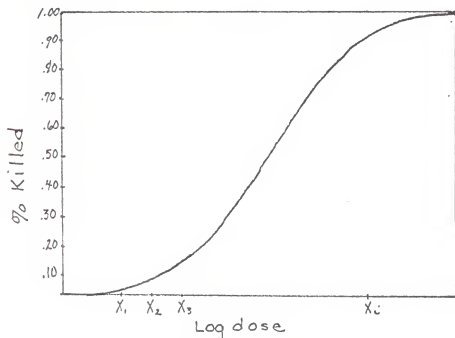


Fig. 2. Ordinates give the percentage of organisms in a sample of size n responding to a dose less than or equal to X .

The original assumption of the normality of individual susceptibility has been tested by reversing the above argument. An expected dosage corresponding to every observed dosage obtained experimentally may be determined from the

fitted curve. These then might be plotted, as was originally impossible, and the normal frequency distribution obtained.

In order to rectify the curve, the standard deviation corresponding to any observed mortality rate may be read directly from the Kelly-Wood Table or the Shepard-Galton Table. When standard deviations are used all the observations below fifty percent kill would have negative expected dosages, which are not convenient. In order to avoid this difficulty, R. A. Fisher devised the probit which is equal to $\lambda + 5$ where λ is the standard normal variate. The expected dosages can be expressed in terms of probits without changing our basic assumptions. The probits corresponding to each percentage killed have been tabled. The assumptions used in Fisher's probits may be summarized as follows:

1. Probability of death of an organism in any experiment is equal to P and is determined by the tolerance of the organism. The probability that r organisms respond out of n tested is $\binom{n}{r} p^r q^{n-r}$.
2. Transforming percentage kill to probits, the probability that an organism is killed is given by equation (3), where Y is the probit or normalized dose plus 5;

$$Y = \frac{X - \mu}{\sigma} + 5 = \alpha + \beta X$$

Therefore:

$$(3) \quad P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Y-5} e^{-\frac{1}{2}\lambda^2} d\lambda.$$

Equation (3) gives:

$$P = P_{\lambda} [\lambda \leq Y - 5] = P_{\lambda} [\lambda + 5 \leq Y]$$

Therefore probit (Y) = $\lambda + 5$ where λ is distributed as a standard normal variable.

If the value of the transform is plotted against the log dose \underline{x} , the

resulting graph is a straight line whose slope and intercept estimate the parameters of the original function, i.e.: $\beta = \frac{1}{\sigma}$, $\alpha = 5 - u/\sigma$.

The first step in probit analysis is to transform each percentage kill to its probit value. If the probit values are plotted as ordinates against some arithmetic function of the amount of dosage (with dosage having equal increments), it is usually not a straight line that is realized but rather a graph which is convex upwards. This deviation from linearity is not entirely unexpected since most dosage-mortality curves are not symmetrical. It was pointed out by Galton in 1879 that variation in biological material follows a geometrical rather than arithmetic distribution, thus suggesting that response might be symmetrical on a log-dose scale. This biological variation has been traced to the relationship between the dose administered and the amount of poison fixed by essential cells or tissues. The use of a log function of dosage produces a symmetrical sigmoidal curve, and also a successful rectification of it, for many different situations. Even though the tolerances of individual units may vary geometrically, it could be considered probable that the average susceptibilities of populations of single cells are normally distributed. Each organism could then be considered as an average of its component cells so that individual organisms may be expected to respond normally to a specific poison.

To begin the analysis, the probit value for each percentage killed is plotted against the corresponding log dose and a provisional regression line is determined. This first estimate of the transformed curve is ordinarily not calculated, but drawn in free-hand. If the observed values are quite scattered, however, the experimenter may choose to calculate the slope of the linear regression line from:

(4)

$$b = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$$

where Y = probit, and X = log-dose. The provisional regression serves two purposes:

1. It determines the probit values for and mortality rates of zero and one hundred percent. Mortality rates of zero and one hundred percent cannot be tabled since the curve of the normal distributions approaches $-\infty$ for zero percent kill and $+\infty$ for one hundred percent kill. This extension of the provisional regression line is acceptable with large sample sizes, but breaks down when the sample size is small. R. A. Fisher showed that when zero survivors are observed the expected probit term for one hundred percent kill is always less than it would be if the class of zero survivors could exert its proper influence on the provisional regression line. (Annals of Applied Biology, 1935). Fisher has supplied a table of corrections which are added to the expected probit given by the provisional regression line. The corrected probit is used for one hundred percent kill.
2. It specifies the appropriate weights to be given to the separate observations in the series. In order to weigh more heavily those observations which are the most reliable, the weights used will be the reciprocals of the variances. The variance needed is that of the probit corresponding to a single observed percentage mortality. The variance of a probit is equivalent to the variance of a percentile (Bliss, 1935). The formula for the variance of a percentile is given by Kelley (cited in Bliss, 1935) as: $\frac{\sigma^2 p_0}{Z^2 n}$ where Z is the ordinate of the normal curve for a given probit, σ

is the standard deviation, and P, Q, and n, have their previous significance. The weight for each observation simplifies to:

$$(5) \quad nw = \frac{nZ^2}{PQ} ,$$

because the probit is already in terms of the standard deviation, i.e. $\sigma^2 = 1$. The term Z^2/PQ is called the weighting coefficient and has been tabled for each 0.1 probit within the useful range of probit values (Bliss, 1935).

A new regression line can now be calculated by the method of maximum likelihood. Because probits and the weights must be estimated from the data, and the weights involve the quantity ultimately to be estimated, namely P, the true percentage dead for each log-dose X, the solution of the maximum likelihood equations must depend upon an iterative process. Adjustments to the values obtained from the provisional regression line are calculated from first order Taylor expansions. The improved values are used as a basis for the second cycle of calculation, and its action continues until adequate convergences to the solutions is reached.

The likelihood function (L) is defined as:

$$(6) \quad L = \sum_{i=1}^k \binom{n_i}{r_i} p^{r_i} q^{n_i - r_i}$$

where $i = 1, 2, \dots, k$ is the index for the different dosages used. The log of the likelihood is given by:

$$(7) \quad \text{Log } L = \sum_{i=1}^k \left[\log \binom{n_i}{r_i} + r_i \log p_i + (n_i - r_i) \log q_i \right].$$

$$\text{Let } S = \log(\bar{x}) + r \log P + (n-r) \log q$$

The maximum likelihood estimates for α and β are found by solving the following equations:

$$(8) \quad \frac{\partial \log L}{\partial \alpha} = \left\langle \frac{\partial S}{\partial P} \cdot \frac{\partial P}{\partial Y} \cdot \frac{\partial Y}{\partial \alpha} \right\rangle \stackrel{\text{set}}{=} 0$$

$$(9) \quad \frac{\partial \log L}{\partial \beta} = \left\langle \frac{\partial S}{\partial P} \cdot \frac{\partial P}{\partial Y} \cdot \frac{\partial Y}{\partial \beta} \right\rangle \stackrel{\text{set}}{=} 0.$$

Evaluation of the partial derivatives gives:

$$(10) \quad \frac{\partial \log L}{\partial \alpha} = f_1(\alpha; \beta) = \left\langle \frac{Zn(p-\hat{P})}{\hat{P}\hat{Q}} \right\rangle = 0$$

$$(11) \quad \frac{\partial \log L}{\partial \beta} = f_2(\alpha, \beta) = \left\langle \frac{Zn(p-\hat{P})(X-\bar{x})}{\hat{P}\hat{Q}} \right\rangle = 0.$$

Because equations (10), (11) cannot be solved as they are, $f_1(\alpha, \beta)$ and $f_2(\alpha, \beta)$ are expressed in terms of a Taylor expansion,

$$f(\alpha, \beta) = f(\alpha_0, \beta_0) + \left. \frac{\partial f}{\partial \alpha} \right|_{\alpha_0, \beta_0} \Delta \alpha + \left. \frac{\partial f}{\partial \beta} \right|_{\alpha_0, \beta_0} \Delta \beta + h(\Delta),$$

where

$$\Delta \alpha = \alpha - \alpha_0.$$

$$\Delta \beta = \beta - \beta_0.$$

The term $h(\Delta)$ involves higher powers of $\Delta \alpha$ and $\Delta \beta$ and will be neglected.

(A detailed solution for the estimates of α and β from the Taylor expansion for $f_1(\alpha, \beta)$ and $f_2(\alpha, \beta)$ can be found in Gilliland, 1964.)

Berkson (1946) gives the approximation

$$p - \hat{P} \doteq Z(Y - \hat{Y})$$

to express the equations, to be solved, in terms of probits rather than percentage mortality; i.e.,

$$(12) \quad \frac{Zn(p-\hat{P})}{\hat{P}\hat{Q}} \doteq \frac{Z^2n(Y-\hat{Y})}{\hat{P}\hat{Q}}$$

and

$$(13) \quad \frac{Zn(p-\hat{P})(X-\bar{x})}{\hat{P}\hat{Q}} \doteq \frac{Z^2n(Y-\hat{Y})(X-\bar{x})}{\hat{P}\hat{Q}}.$$

The solutions (Gilliland, 1964) give the weighted standard regression

equations in probits (Y) and log-dose (X). Using w for the weight Z^2/PQ , the following equations are used to estimate β and α :

$$(14) \quad b = \text{estimate of } \beta = \frac{\sum n w (X - \bar{X})(Y^* - \bar{Y}^*)}{\sum n w (X - \bar{X})^2}$$

$$(15) \quad a = \text{estimate of } \alpha = \bar{Y}^* - b\bar{X}.$$

The probits Y are obtained from the provisional regression line and \bar{X} and \bar{Y} are the weighted means of the log-dose and probit values respectively

$$\bar{X} = \frac{\sum n w X}{\sum n w}$$

$$\bar{Y}^* = \frac{\sum n w Y^*}{\sum n w}$$

In terms of the original normal distribution

$$b = \text{estimate of } 1/\sigma$$

$$a = \text{estimate of } 5 - \mu/\sigma$$

In summary, the calculations for estimating α and β may be carried out as follows:

1. Find the provisional regression line. From each observed percentage dead obtain the corresponding probit value Y from the tables and plot Y against log-dose X. A straight line fitted by eye is then used to obtain a set of expected probits, \hat{Y} corresponding to log-dose X.
2. To obtain a second approximation to the regression line, a series of working probits and their weights are obtained, which are then used in the regression formulas previously given. The working probits are obtained from either:

$$Y_1^* = \hat{Y} + \hat{Q}/Z - q/z$$

or

$$Y_1^* = \hat{Y} - \hat{P}/Z + p/z$$

where \hat{Y} is the expected probit obtained from the previous regression line, p = observed percentage of individuals responding, and $q = 1 - p$. The appropriate weighting coefficient for Y_1 is obtained from a table. Also tabled along with the values of the weights for the probit value are the maximum working probit $Y_{\max} = Y + Q/Z$, the minimum working probit $Y_{\min} = Y - P/Z$, and the Range $R = 1/Z$. Y_1 and nw_1 are then used in the regression formulas to compute the new estimates of the regression coefficients α and β .

3. The values of the expected probits obtained from this second approximation to the regression line may then be used to repeat the iterative process. Iteration is continued until the desired degree of accuracy is reached. It might be noted here that from a statistical point of view b is the slope with which the regression line passes through the point (\bar{X}, \bar{Y}) ; while from a biological point of view, b measures how closely the individual organisms in the experiment agree with one another in their sensitivity to the toxic agent. If a small change in dosage concentration gives a wide range in the percentage kill, the sensitivity is high. This toxicological characteristic can be expressed as the percentage increase in dosage that is required to increase kill by one probit. This is given by the ratio:

$$(17) \quad \frac{100 \log_e 10}{b} = \frac{230.6}{b} \quad (\text{Bliss, 1935})$$

The most likely position of the true dosage-mortality curve for the entire population has been computed on the basis of the experimental evidence obtained from a sample. A different sample would have produced a different regression line. To determine how accurately the curve has been determined, i.e., whether the observed mortalities agree with the theoretical mortalities obtained from the regression line, a chi-square test may be used. If none of the expected frequencies nP or nQ is too small (less than 5) the formula for Pearson's chi-square may be obtained from the following table:

TABLE I

	DEAD	ALIVE	TOTAL
OBSERVED	pn	qn	n
EXPECTED	Pn	Qn	n

Number of organisms observed to be dead or alive out of a sample of size n.

χ^2 then is computed to be:

$$\begin{aligned}
 \chi^2 &= \sum \frac{(pn - \hat{P}n)^2}{\hat{P}n} + \frac{(qn - \hat{Q}n)^2}{\hat{Q}n} \\
 (18) \quad &= \sum \frac{n}{\hat{P}\hat{Q}} (p - \hat{P})^2 \\
 &= \sum \frac{(r - n\hat{P})^2}{n\hat{P}\hat{Q}}
 \end{aligned}$$

where r = number responding to dosage X . An easier method of computation given by Bliss (1935) adapted from one given by Fisher, is as follows:

$$(19) \quad \chi^2 = [LwY^2 - \bar{Y} LwY] - b [LwXY - \bar{X} LwY].$$

Nearly all of these components were computed in determining the regression equation. The number of degrees of freedom is two less than the number of

levels of X used in the experiment.

In dealing with expected frequencies less than 5, two methods have been proposed:

1. The exact procedure, according to Bliss (1935), would be to exclude from the computation of χ^2 the results of those dosages at which the expected survivors (or mortalities) are less than 5.
2. The second method is to group together the results of those dosages in which the expected survival rate is small (or expected death rate is small), since they will contribute less to the χ^2 value as a group, than as single observations.

If the calculated χ^2 with $(n - 2)$ d.f., is significant, then either the observations depart significantly from a straight line relationship, or some uncontrolled condition in the experiment is causing a greater variation about the line than can be attributed to fluctuations due to sampling. According to Bliss (1935) the latter is the more likely, since systematic deviations from linearity were eliminated from the start.

It is useful to see how accurately a and b have been estimated. The formulas for the variances of a and b, as given by Bliss (1935), are:

$$\text{Var } (b) = S^2b = \frac{\chi^2}{n[\sum wX^2 - \bar{x}\sum wX]}$$

$$\text{Var } (a) = S^2a = \frac{\chi^2}{n\sum w}$$

When the χ^2 test for the position of the computed curve is non-significant, these variances may be reduced to a simpler form for all tests involving the same dosages and numbers of organisms. Replacing χ^2/n by its approximate expected value when n is large,

$$E [X^2/n] = \frac{n-2}{n} \doteq 1,$$

the variances simplify to:

$$\text{Var (b)} = \frac{1}{\sum wX^2 - \bar{X}\sum wX}$$

$$\text{Var (a)} = \frac{1}{\sum w}.$$

Confidence limits can also be placed about the regression line, as given by:

$$Y = a + b (X - \bar{X}) \pm t_{n-2} [S^2a + (X - \bar{X})^2 S^2b]^{\frac{1}{2}}$$

where t is taken from "Students'" table for the appropriate α - level of confidence and $(n - 2)$ degrees of freedom (Bliss, 1935).

It was given previously that \underline{a} was the estimate of $5 - \mu/\sigma$. Therefore the estimate of μ is given by $(5 - \underline{a})/b$ where b is the estimate of $1/\sigma$. This is the estimate of the dosage at which there is a fifty percent response, often called the L. D. 50 (median lethal dose). The large sample formula for the variance of X_{50} (estimate of L. D. 50) is given by:

$$\sigma_{X_{50}}^2 = \frac{1}{\sum w} + \frac{(X_{50} - \bar{X})^2}{\sum w (X - \bar{X})^2}$$

(cited in Biometrical Tables for Statisticians as given by Finney, 1952)

3. LOGIT ANALYSIS

Logit analysis was devised to avoid the necessity of assuming normality of tolerances. According to Berkson (1951), the purpose of the analysis is to give information about the relationship between dosage and response given by the slope of the dosage-mortality curve, not deviations of hypothetical

tolerances of the organisms. When speaking of the standard deviation of tolerances, a variability is perhaps being supplied for something that in fact does not exist at all. Berkson offered several examples to question the existence of tolerances. One involved subjecting pilots to high-altitude conditions and observing their all-or-none responses of getting the "bends". Those who succumbed were marked as having low tolerance and were not to be assigned to high-altitude flying. At Berkson's request the tests were repeated and many of the pilots who had succumbed the first time were not affected the second time; and many of those who passed the first trial, got the "bends" on the second. Either their tolerances had changed; or as he questions "Did it exist at all?". A second example was a bio-assay experiment in which *Drosophila* were exposed to increasing doses of X-ray intensity and the quantal response of the percentage mutation was observed. From the probitist's point of view, the increase in percentage response from one dose to the next reflected a difference in the tolerances of the flies. However, it is the generally accepted theory that the probability of an effective hit on a cell by a photon, which is what causes the mutation, is directly proportional to the intensity of the radiations. In other words the more photons the more chance of mutation, rather than the cell of a fly showing a tolerance to photons.

If assumptions of normality are discarded and with it the use of the integrated normal in fitting the dosage-mortality curve, a new functional relationship must be found to govern the process. Berkson, on the basis of previous work done in studies of population growth, has suggested the logistic curve:

$$(26) \quad Y = \frac{1}{1 + e^{-(\alpha + \beta X)}}$$

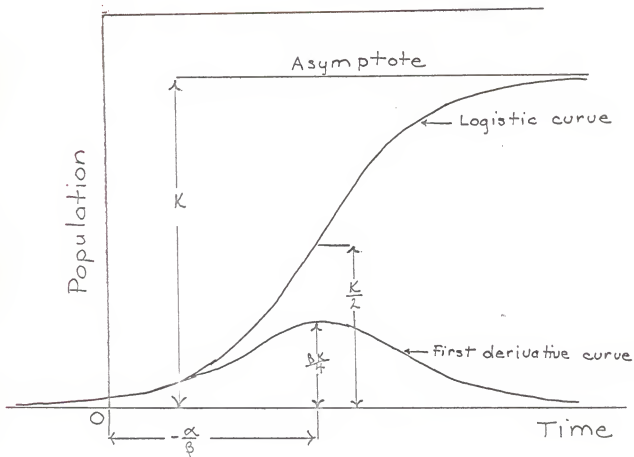


Fig. 3 The logistic curve and its first derivative curve as depicting population growth over time.

3.1 THE LOGISTIC CURVE

In 1838 P. F. Verhulst, a Belgian mathematician, suggested the use of a curve which he called the "logistic" to describe the growth of human populations. His work was forgotten for many years and in 1920 R. Pearl and L. J. Reed, without knowing of Verhulst's work, derived the logistic curve empirically to meet certain postulates for a curve to describe population growth. The equation for the logistic may be written in the form:

$$(27) \quad Y = \frac{K}{1 + e^{-(\alpha + \beta t)}}$$

The first derivative with respect to time gives the change in mass per unit of time:

$$\frac{dY}{dt} = \frac{K\beta e^{-(\alpha + \beta t)}}{[1 + e^{-(\alpha + \beta t)}]^2},$$

which can be rewritten as follows:

$$\begin{aligned} \frac{dY}{dt} &= \frac{K}{[1 + e^{-(\alpha + \beta t)}]} \cdot \frac{\beta e^{-(\alpha + \beta t)}}{1 + e^{-(\alpha + \beta t)}} \cdot \frac{K}{K} \\ &= \frac{\beta Y [K + e^{-(\alpha + \beta t)} K - K]}{[1 + e^{-(\alpha + \beta t)}] K} \\ &= \beta Y \left[\frac{K (1 + e^{-(\alpha + \beta t)})}{K (1 + e^{-(\alpha + \beta t)})} - \frac{K}{(1 + e^{-(\alpha + \beta t)}) K} \right] \\ &= \beta Y \left[1 - \frac{Y}{K} \right] \end{aligned}$$

or
 (28)
$$\frac{dY}{dt} = \frac{\beta Y (K - Y)}{K} .$$

Using the above equations and Fig. 2 the following properties of the logistic are derived:

1. The logistic is asymptotic to a line K units above the t -axis and parallel to it.
2. The point of inflection is given by the coordinates:

$$t = -\alpha/\beta \text{ and } Y = K/2.$$

This is shown as follows:

$$\begin{aligned} \frac{d^2Y}{dt^2} &= \frac{\beta}{K} \left[Y \left(-\frac{dY}{dt} \right) + (K - Y) \frac{dY}{dt} \right] \\ &= \frac{\beta}{K} \left[K \frac{dY}{dt} - 2Y \frac{dY}{dt} \right] . \end{aligned}$$

Setting $\frac{d^2Y}{dt^2} = 0$;

$$K \frac{dY}{dt} - 2Y \frac{dY}{dt} = 0$$

$$K - 2Y = 0$$

(29)

$$Y = K/2.$$

Solving for t and using $Y = K/2$:

$$K/2 = \frac{K}{1 + e^{-(\alpha+\beta t)}}$$

$$1/2 = \frac{1}{1 + e^{-(\alpha+\beta t)}}$$

$$1 + e^{-(\alpha+\beta t)} = 2$$

$$e^{-(\alpha+\beta t)} = 1$$

hence

$$\frac{dY}{dt} e^{-(\alpha+\beta t)} = 0$$

$$-(\alpha+\beta t) = 0$$

so that the point of inflection is given by

$$(30) \quad t = -\alpha/\beta, \quad Y = K/2$$

3. The rate of change of the mass Y is greatest at $Y = K/2$ and is given by:

$$(31) \quad \left. \frac{dY}{dt} \right|_{Y=K/2} = \frac{\beta(K/2)(K-K/2)}{K} = \frac{\beta K}{4}.$$

4. By inspection of equation (28)

$$\frac{dY}{dt} = \frac{\beta Y(K-Y)}{K},$$

the rate of growth of the population, diminishes with time as a result of the slowing effect of the factor $(K - Y)$, which measures the aggregate of forces that slow down and finally stop growth.

Because of the dynamic relationship expressed in equation (28) above, where the rate of change of the mass Y with respect to time t is proportional to a factor that decreases as Y increases, the logistic has been successfully applied in a great many experimental fields. It has been used to describe chemical autocatalysis, electrode potential of an oxidation-reduction reaction, enzyme reactions, and other organic reactions, as well as the previously stated population growth and the growth of an individual. Thus, the logistic function applies to a wide range of phenomena whose physical mechanisms are different. And all of these mechanisms are dynamic, as opposed to a static distribution of tolerances.

While the logistic is not considered as a probability density function in logit analysis it is of additional statistical interest to note that the

density function of the logistic has been given by Gupta (1965).

A random variable Y is said to follow a logistic distribution $L(\mu, \sigma^2)$ if its cumulative distribution function is:

$$F(Y; \mu, \sigma) = \frac{1}{1 + \exp\left\{-\frac{(Y-\mu)}{\sigma} \cdot \frac{\pi}{\sqrt{3}}\right\}}$$

The probability density function is:

$$f(Y; \mu, \sigma) = \frac{(\pi/\sqrt{3}) \sigma \exp\{-\pi(Y-\mu)/\sqrt{3} \sigma\}}{[1 + \exp\{-\pi(Y-\mu)/\sqrt{3} \sigma\}]^2}$$

where

$$-\infty < Y < \infty$$

and

$$-\infty < \mu < \infty, \sigma > 0.$$

The density is symmetrical with mean μ and variance σ^2 . The moment generating function of $X = \frac{(Y-\mu)}{\sigma}$ is:

$$M_x(t) = (1+t/g) (1 - t/g)$$

where $g = \pi/\sqrt{3}$.

3.2 APPLYING THE LOGISTIC TO BIOASSAY

In logit analysis, the logistic function is used to describe the true mortality rate P instead of a normal distribution of tolerances:

$$(35) \quad P = \frac{1}{1 + e^{-(\alpha + \beta t)}}$$

Comparing the dosage-mortality curve of a bioassay experiment (Fig. 4) with the population growth curve graph used by Pearle and Reed, the following facts

may be noted:

1. The population growth has been replaced by the death rate, thus where K stood for the maximum population level in the Pearle and Reed model K is now equal to one.
2. The time intervals on the abscissa have been replaced by the increasing concentrations of the doses administered.

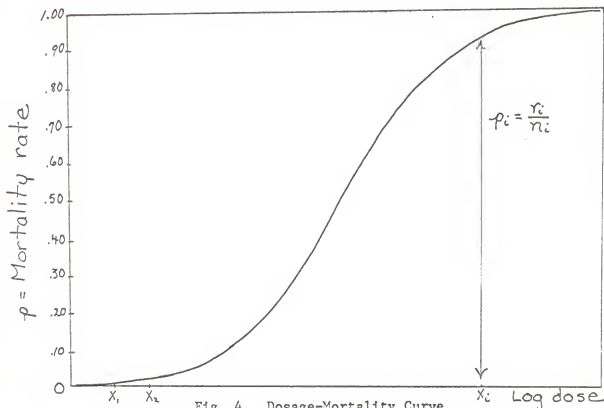


Fig. 4. Dosage-Mortality Curve.

The probability of observing r deaths out of n organisms exposed to a dose X is again governed by the binomial distribution and is given by:

$$(36) \quad \binom{n}{r} p^r q^{n-r}$$

In fitting the logistic to population growth, Pearl and Reed used the method of least squares. The weighted normal equations are obtained by minimizing the quantity:

$$(37) \quad \sum \frac{n}{\hat{P}\hat{Q}} (p - \hat{P})^2$$

(where the weight is the reciprocal of the variance of P .) This cannot be solved directly in terms of the logistic because:

1. The logistic is not linear in the parameters to be evaluated.
2. The weights contain the quantities P and Q to be estimated. The logistic can be expanded in terms of a Taylor series and solution obtained by successive approximations, as was done with probits.

To avoid the first difficulty to an easy solution, a transformation is made which will rectify the curve in the same manner as used in probit analysis.

The transformation used is the logit, defined as:

$$\begin{aligned} \text{Logit } L &= \ln \frac{P}{1-P} \\ &= \ln \frac{1}{1 + e^{-(\alpha + \beta X)}} \\ &= \ln \frac{1}{1 + e^{-(\alpha + \beta X)}} \\ L &= \ln \frac{1}{e^{-(\alpha + \beta X)}} \end{aligned}$$

$$(38) \quad = \alpha + \beta X.$$

The logit L can then be plotted against the corresponding dosage X and the resulting straight line gives the original parameters as the slope and intercept.

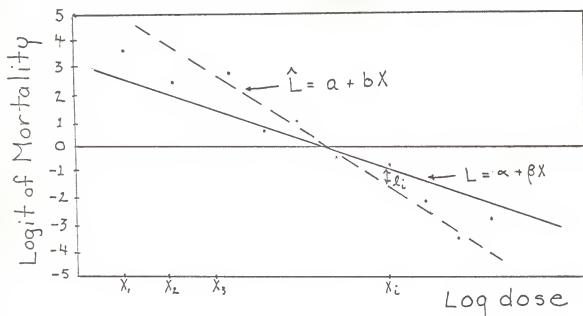


Fig. 5. Scatter diagram showing true regression line and estimated regression line.

The L. D. 50 is the dosage such that fifty percent of the organisms exposed die, that is:

$$P = 1/2 = \frac{1}{1 + e^{-(\alpha + \beta X_{50})}}$$

Using elementary algebra one obtains:

$$1 = e^{-(\alpha + \beta X_{50})};$$

and taking natural logarithms;

$$0 = -(\alpha + \beta X_{50})$$

giving the dose L. D. 50 = $-\alpha/\beta$.

Making an approximation analogous to that used in the solution of the maximum likelihood equations for probits (Berkson, 1946), the minimum X^2 can be found in terms of the logit $l = \ln(p/q)$ rather than the observed response p .

Therefore (39) $(p - \hat{P})^2 \doteq (\hat{PQ})^2 (l - \hat{L})^2;$

or, using a further approximation

(40) $(p - \hat{P})^2 \doteq (\hat{PQ})(pq) (l - \hat{L})^2.$

Using equation (18)

$$(40) \quad \mathcal{X}^2 = \sum \frac{n}{pq} (p - \hat{p})^2$$

We are now able to determine estimates for α and β by minimizing the logit \mathcal{X}^2 .

$$(41) \quad \mathcal{X}^2 = npq (\hat{L} - \hat{L})^2$$

(Berkson originally called this the method of least squares.)

An iterative solution for logits analagous to that used for probits could be used, but by minimizing equation (41) a direct solution can be obtained since the weights on the right side are entirely in terms of the observed values, n , p , q ; i.e., they do not contain the parameters to be estimated. This same simplification does not occur with the integrated normal since the corresponding approximation would be

$$(42) \quad (p - \hat{p})^2 \div \hat{Z}^2 (Y - \hat{Y})^2$$

where y is the observed probit, \hat{Y} is the expected probit, and \hat{Z} is the estimated normal ordinate. Using this in the same \mathcal{X}^2 (18) we do not divide out the $\hat{P} \hat{Q}$ in the denominator.

Berkson gives the following properties for all values of the parameters of the minimum logit \mathcal{X}^2 (Berkson, 1955)

1. The logit \mathcal{X}^2 is distributed asymptotically as \mathcal{X}^2 .
2. It is asymptotically efficient.
3. It is sufficient.
4. It has a smaller sampling error (mean square error) and smaller variance about the mean than the maximum likelihood estimate.

To obtain the normal equations for estimating α and β , the logit \mathcal{X}^2 is differentiated with respect to α , and with respect to β . As with probits the derivative must be composite since the logit $L = \alpha + \beta x$.

$$(43) \quad \frac{\partial X^2}{\partial \alpha} = \frac{\partial X^2}{\partial L} \cdot \frac{\partial L}{\partial \alpha} = \sum npq (l - \hat{L}) \stackrel{\text{set}}{=} 0$$

$$(44) \quad \frac{\partial X^2}{\partial \beta} = \frac{\partial X^2}{\partial L} \cdot \frac{\partial L}{\partial \beta} = \sum npqx (l - \hat{L}) \stackrel{\text{set}}{=} 0$$

The solution of equations (43) and (44) lead to the least squares solutions for the regression line $\hat{L} = a + bX$, where \underline{npq} is the weight of dose X . The estimates of α and β are given by

$$(45) \quad b = \frac{\sum npq (l - \bar{l})(X - \bar{X})}{\sum npq (X - \bar{X})^2}$$

$$(46) \quad a = \bar{l} - b\bar{x}. \quad (\text{Berkson, 1953}) ,$$

where

$$(47) \quad \bar{L} = \frac{\sum npql}{\sum npq}$$

$$(48) \quad \bar{x} = \frac{\sum npqX}{\sum npq}$$

The weights $w = pq$ and $w1 = pql$ have been tabulated (Berkson, 1953) using the machine formula:

$$(49) \quad b = \frac{\sum nw1X - \frac{\sum nw1 \sum nwX}{\sum nw}}{\sum nwX^2 - \frac{(\sum nwX)^2}{\sum nw}}$$

$$(50) \quad a = \frac{\sum nw1 - b \sum nwX}{\sum nw}$$

The estimate of the L. D. 50 is given by $X_{50} = -a/b$.

A close approximation of a least squares solution in terms of the logistic has now been obtained with only the arithmetic of the estimates involved.

In the instances in which the observed mortality rate is zero or 100 percent the logit cannot be used since it becomes infinite at these values. The

method used for probits can be used to obtain a preliminary solution for the observations in question.

The variances of a and b , the estimates of α and β , have been derived as the asymptotic variances with estimates then substituted for the parameters (Berkson, 1958). The estimated logit linear equation may be written as:

$$\hat{L} = a + bX = a' + b(x - \bar{x})$$

where $a' = \bar{L} = a + b\bar{x}$. The formulas for the variances of the estimates of the parameters may be written as follows:

$$(51) \quad S_{a'}^2 = \frac{1}{\sum w}$$

$$(52) \quad S_b^2 = \frac{1}{\sum w(x - \bar{x})^2}$$

$$(53) \quad S_a^2 = S_{a'}^2 + \bar{x}^2 S_b^2 = \frac{1}{\sum w} + \frac{\bar{x}^2}{\sum w(x - \bar{x})^2}$$

These formulas provide closely accurate estimates of the variances, under the ideal conditions in which:

1. The true P 's are given exactly by equation (35).
2. The samples are random for each fixed dose.
3. The number of organisms used at each dose is large (Berkson, 1953).

The variance of the estimate of the L_{50} , where X is the log-dose, is given by

$$(54) \quad S_{x_{50}}^2 = \frac{1}{b^2} S_a^2 + S_b^2 (x_{50} - \bar{x})^2$$

The use of logit analysis leads to a relatively easy solution for the estimates of the parameters, and thus an easily fit dosage-mortality curve. From the mortality rate p at any log-dose X the logits \hat{L} and weights w may be obtained from tables. (Berkson, 1953) The antilogits, p , for logit \hat{L} have

also been tabulated by Berkson. The straight line transform obtained in terms of a and b may then be easily plotted using special logit graph paper, sold by the Codex Book Company, Norwood, Massachusetts. From the regression line

$$\hat{L} = a + bx$$

the expected logits may be found. Using the antilogit tables (Berkson, 1955) to find p and Q for logit L , the Pearson chi-square

$$(55) \quad \chi^2 = \sum \frac{n(p - \hat{p})^2}{\hat{p}\hat{q}}$$

can be calculated. The logit chi-square could have been calculated instead of the Pearson chi-square, which as was indicated earlier, reduces the computational work.

The accuracy of the estimates for α and β can be measured by the formulas for the variance of a and the variance of b .

Because the assumption of normality has been discarded in favor of the logistic distribution in logit analysis, the statistical tests used in probit analysis cannot be used.

4. CONCLUSION

The integrated normal curve used in probit analysis and the logistic curve used in logit analysis lead to essentially the same rectification of the original curvilinear data and to essentially the same fitted straight line. In practice, discrimination between the normal and logistic fitting of the dosage-response relationship is not likely to be possible. When the chi-square values for probits and logits are compared, the results are practically the same (Berkson, Finney). Berkson (1944, 1946) gives data which indicates that:

1. Both the probit and logit chi-square approach the true value of

chi-square.

2. The logit approximation is closer than that of probits.
3. The logit approximation always gives a smaller value than the true value, while the probit approximation is sometimes lower, sometimes higher.
4. The variability from the true value is greater for the probit than for the logit.

While the final fitted curve is nearly the same for both methods, the calculations involved are definitely more laborious in probit analysis than logit analysis. From the purely empirical curve-fitting point of view, logit analysis might be preferred.

If more than curve-fitting is desired in the analysis, then the assumptions behind probits or logits should be considered. Probit analysis assumes a normal distribution of tolerances of the organisms to the agent. Logit analysis was devised to avoid such a static distribution of tolerances. But, by not making the assumption of normality, no tests or confidence intervals about the estimates can be run. It must be recognized that the true distribution may not be normal, but in the absence of evidence favoring a specific alternative, the hypothesis of normality is very attractive (Finney, 1952). In fact, the central limit theorem gives reason for hoping that conclusions based on the normal assumption will be close to the truth if the means of several observations are involved.

The differences involved in using logit analysis or probit analysis have aroused considerable interest and some controversy in bioassay. Perhaps it should be recognized that both probit and logit analysis are important in bioassay. The choice between them may depend upon the nature of the biological reactions in use, and the results desired.

REFERENCES

- Berkson, Joseph, M.D. (1944). Application of the logistic function to bio-assay. J. Amer. Statist. Assoc. 39: 357-365.
- Berkson, Joseph, M.D. (1946). Approximation of chi-square by "probits" and by "logits". J. Amer. Statist. Assoc. 41: 70-74.
- Berkson, Joseph, M.D. (1949). Minimum X^2 and Maximum likelihood solution in terms of a linear transform, with particular reference to bio-assay. J. Amer. Statist. Assoc. 44: 273-278.
- Berkson, Joseph, M.D. (1951). Why I prefer logits to probits. J. Amer. Statist. Assoc. 7: 327-339.
- Berkson, Joseph, M.D. (1953). A Statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. J. Amer. Statist. Assoc. 48: 565-599.
- Berkson, Joseph, M.D. (1955). Maximum likelihood and minimum X^2 estimates of the logistic function. J. Amer. Statist. Assoc. 50: 130-162.
- Bliss, C. I. (1935). The calculation of the dosage-mortality curve. Annals of Applied Biology. 22: 134-167.
- Bliss, C. I. (1935). The comparison of dosage-mortality data. Annals of Applied Biology. 22:307-333.
- Bliss, C. I. (1952). The Statistics of Bioassay. Academic Press Inc., New York.
- Cramer, Harold. (1946). Mathematical methods of statistics. Princeton University Press, Princeton, New Jersey.
- Dixon, W. J. (1965). The up-and-down method for small samples. J. Amer. Statist. Assoc. 60: 967-978.
- Finney, D. J. (1952). Statistical Method in Biological Assay. Hafner Publishing Co., New York.

- Fryer, H. C. (1966). Concepts and Methods of Experimental Statistics.
467-478. Allyn and Bacon, Inc., Boston, Mass.
- Gilliland, P. D. (1964). Some Mathematical Aspects of Probit Analysis, A
Master's Report, Kansas State University.
- Gupta, S. S. and Shah, B. K. (1965). Exact moments and percentage points of
the order statistics and the distribution of the range from the logistic
distribution. Ann. Math. Stat. 36: 907-920.
- Hodges, J. L., Jr. (1958). Fitting the logistic by maximum likelihood.
J. Biometric Soc. 14: 453-461.
- Linder, Arthur (1964). Statistics of Bioassay. Institute of Statistics Mimeo
Series No. 404.
- Pearl, R. (1940). Introduction to medical biometry and statistics. W. B.
Saunders Company, Philadelphia, Pa.
- Pearson, E. S. and Hartley, H. O. (1962). Biometrika tables for statisticians.
1: 4-9. The Syndics of the Cambridge University Press, London, England.
- Wilson, E. B. and Worcester, J. (1943). The determination of L. D. 50 and
its sampling error in bio-assay, II. National Academy of Sciences.
29: 114-120.
- Wilson, E. B. and Worcester, J. (1943). The determination of L. D. 50 and
its sampling error in bio-assay, III. National Academy of Sciences.
29: 257-262.

ACKNOWLEDGEMENTS

The writer wishes to express her sincere appreciation to her major professor, Dr. Holly C. Fryer, for suggesting the topic of this report and his assistance in its preparation. The writer wishes also to thank her husband, Patrick L. Duncan, for his several typings of the manuscript and general patience in her confusion.

PROBIT AND LOGIT ANALYSES

by

JANET LEE DUNCAN

B.S, Kansas State University, 1963

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Probit and logit analyses are generally utilized in the area of bioassay. In bioassay plants or animals are subjected to some stimulus in varying intensities in order to determine the relationship between the intensities of the stimulus and the responses they elicit from the organism. When the response is known as a function of the stimulus, predictions can be made of intensities which produce desirable responses.

Probit and logit analyses deal with a particular type of bioassay where the response by an organism to a stimulus administered is quantal. Quantal responses are all-or-none responses in which the organism is either affected or not affected by the stimulus, as opposed to responses measured on a continuous scale.

The purpose of this report is to compare the methods of obtaining this dosage-response relationship by probit analysis and logit analysis.

Probit analysis and logit analysis differ basically in the underlying assumptions and in the method of fitting the dosage-response curve.

Probit analysis assumes a normal distribution of tolerances of the organism to the stimulus. The probability that a certain percentage of organisms will respond to a given dose of the stimulus is given by

$$P = \int_{-\infty}^{Y-5} \frac{1}{\sqrt{2\pi}} e^{-\frac{\lambda^2}{2}} d\lambda$$

where Y is the probit or normalized percentage killed, plus five; i.e.,

$$\text{Probit } (Y) = \lambda + 5$$

The sigmoidal curve obtained by plotting the percentage responding against a given dosage is rectified (made linear) by using the probit transformation and plotting probits against dosage. The method of maximum likelihood is used to estimate the parameters of the regression line obtained in terms of probits

and dosage. Because the weighted normal equations contain the parameters to be estimated, a first order Taylor expansion is used as an approximate solution. An iterative procedure must then be used to solve for the estimates to a desired accuracy. The variance of the estimates can be calculated and confidence limits can be placed about the regression line.

Logit analysis was devised to avoid the assumption of normality. The probability that a certain percentage of organisms will respond to a certain dose of the stimulus is given by the logistic curve:

$$P = \frac{1}{1 + e^{-(\alpha + \beta t)}}$$

The rectification of the curve is accomplished by transforming the percentage responding to logits by:

$$\text{Logit } (Y) = \ln \frac{p}{q}$$

where p is the observed percentage responding and $q = 1 - p$. The minimum chi-square method is used to find the estimates of the parameters, α and β . An approximation of weighted percentage responding in terms of logits enables the normal equations to be solved directly and weighted equations for linear regression are obtained. The variances of the estimates can be calculated, but no confidence limits can be set since normality was not assumed.

If the fitting of the dosage-response curve is all that is wanted and the assumption of normality is not obvious from the material being tested, logit analysis gives the easier solutions.

If tests of the estimates are needed, and the assumption of normality of tolerances is reasonable, probit analysis must be used.

Both methods have an importance in bioassay, the choice between them depending on the nature of the biological reactions in use and the results desired.